Theoretical Computer Science 32 (1984) 47–60
North-Holland

LINEAR INDEXED LANGUAGES

J. DUSKE and R. PARCHMANN
Institut für Informatik, Universität Hannover, D-3000 Hannover 1, Fed. Rep. Germany

Communicated by M. Nivat
Received June 1983
Revised January 1984

Abstract. In this paper one characterization of linear indexed languages based on controlling linear context-free grammars with context-free languages and one based on homomorphic images of context-free languages are given. By constructing a generator for the family of linear indexed languages, it is shown that this family is a full principal semi-AFL. Furthermore a Parikh theorem for linear indexed languages is stated which implies that there are indexed languages which are not linear.

1. Introduction

Indexed grammars and languages were introduced by Aho [1] as an extension of context-free grammars and languages. We will call indexed grammars linear, if the right side of each production contains at most one variable. The family of languages generated by these grammars is called the family of linear indexed languages and will be studied in this paper. Linear indexed languages properly contain the context-free languages.

In the context-free case, linear languages have the following characterizations:

(1) A context-free language is linear iff it can be obtained by a linear context-free grammar controlled with a regular set. (This characterization is trivial.)

(2) A context-free language is linear iff it is of the form $\{h_1(w)h_2(w)^R \mid w \in U\}$, where h_1, h_2 are homomorphisms and U is a regular set.

In Section 3 we will show that by replacing 'regular' by 'context-free' in this characterization we will obtain exactly the family of linear indexed languages.

Furthermore, it is known that the family of linear context-free languages is a full principal trio, which is equivalent to a full principal semi-AFL. We will show that this holds for the family of linear indexed languages too. In Section 4 we will construct, starting with the semi-Dyck set D_2 and using the characterizations obtained in Section 3, a generator for this family.

In Section 5 it will be shown that the images of linear indexed languages under Parikh mappings are semilinear sets. This implies that the family of linear indexed languages is a proper subclass of the family of indexed languages.

2. Basic definitions and facts

Classes of languages which can be generated by grammars with a restricted use of productions are investigated extensively in the literature (see, e.g., [9]). An example of such a grammar with a restricted derivation mode is the controlled (linear) grammar given by Khabbaz [7]. Let us first recall this definition in the following form.

Definition 2.1. Let $G=(N, T, P, S)$ be a linear context-free grammar, M be a finite set, μ a function from M onto P and $L \subseteq M^*$ be a language. Then $C=(G, \mu, M, L)$ is called a controlled linear grammar with control function μ and control set L.

If $\mu(m)$ is the production $A \to \alpha$, then we will write $m: A \to \alpha$. Let $p \in P$ be a production and let $m \in M$ with $\mu(m)=p$. Then we will write $u \Rightarrow_m v$, if v can be obtained from u by application of p. This notation extends to $u \Rightarrow_\pi v$, $\pi \in M^*$, in the usual manner. The language $L(C)$ generated by C is the set $L(C)=\{w \mid w \in T^*$ and $S \Rightarrow_\pi w$ with $\pi \in L\}$.

$L(C)$ is a linear context-free language if L is regular. This fact can easily be shown using [3, Lemma 2.1] and closure properties of linear context-free languages. Conversely, every linear context-free language can be given as $L(C)$, where C is a controlled linear grammar with a regular control set L. This means that the linear context-free languages are exactly those languages obtained by controlling linear context-free grammars with regular languages.

Now the question arises, whether it is possible to characterize those languages which can be obtained by controlling linear context-free grammars with context-free languages. This class of languages is denoted by $\mathscr{L}_1$ in [7] and we will show in Section 3 that this class is exactly the class of linear indexed languages.

Aho [1] introduced indexed grammars and languages. We will state these notions in the following form.

Definition 2.2. An indexed grammar is a 5-tuple $G=(V, T, I, P, S)$, where

(1) V, T, I are finite pairwise disjoint sets; the sets of variables, terminals, and indices, respectively,

(2) P is a finite set of pairs (Af, α), $A \in V$, $f \in I \cup \{e\}$, $\alpha \in (VI^* \cup T)^*$, the set of productions. (Af, α) is denoted by $Af \to \alpha$.

(3) $S \in V$, the start variable.

Let $\alpha = u_1 B_1 \beta_1 u_2 B_2 \beta_2 \ldots B_k \beta_k u_{k+1}$ with $u_i \in T^*$ for $i \in [1:k+1]$, and $B_i \in V$, $\beta_i \in I^*$ for $j \in [1:k]$ with $k \geq 0$ be an element of $(VI^* \cup T)^*$, and let $\gamma \in I^*$.

Then we set $\alpha : \gamma = u_1 B_1 \beta_1 \gamma u_2 B_2 \beta_2 \gamma \ldots B_k \beta_k \gamma u_{k+1}$.

For $u, v \in (VI^* \cup T)^*$ we set $u \Rightarrow v$ iff $u = \phi_1 Af \gamma \phi_2$, $v = \phi_1(\alpha : \gamma)\phi_2$ with $\phi_1, \phi_2 \in (VI^* \cup T)^*$ and $Af \to \alpha \in P$. $\Rightarrow^n$ is the n-fold product and $\Rightarrow^*$ is the reflexive, transitive closure of $\Rightarrow$. The language $L(G)$ generated by an indexed grammar $G=(V, T, I, P, S)$ is the set $L(G)=\{w \mid w \in T^*$ and $S \Rightarrow^* w\}$.

A language L is called an indexed language iff $L = L(G)$ for an indexed grammar G.

Definition 2.3. An indexed grammar $G = (V, T, I, P, S)$ is called linear, if the right side of each production contains at most one variable. A language L is called linear indexed iff $L = L(G)$ for a linear indexed grammar.

The grammar given in [1, Example 1], which generates the indexed language $\{a^n b^n c^n \mid n \geq 1\}$, is a linear indexed grammar. Let us give another example.

Example 2.4. Let $G = (\{S, A, A'\}, \{a, b\}, \{f, g, \#\}, P, S)$ with

$$P = \{S \to A\#,\ A \to aAf,\ A \to bAg,\ A \to A',\ A'f \to A'a,\ A'g \to A'b,\ A'\# \to e\}.$$

This grammar generates the language $L(G) = \{ww \mid w \in \{a, b\}^*\}$.

It is easy to see that the family of linear indexed languages is exactly the family of languages generated by RIL-grammars introduced by Aho [1] (for a proof, see Parchmann [8]). This implies that the family of context-free languages is a proper subclass of the family of linear indexed languages.

Recall that a family of languages is called a full trio if it is closed under homomorphisms, inverse homomorphisms, and intersections with regular sets. We have the following theorem.

Theorem 2.5. *The family of linear indexed languages is a full trio.*

Proof. A grammar-based proof of this theorem can be given in analogy to the linear context-free case (see, e.g., [6, p. 283]). □

Now we want to show that the class of linear indexed languages is not closed under product and Kleene closure.

To this end recall (see [1]) that an indexed grammar $G = (V, T, I, P, S)$ is called RIR-grammar (rightlinear indexed rightlinear) if all productions in P are of one of the forms $Af \to aB$, $Af \to a$ or $A \to aBf$, where $A, B \in V$, $a \in T \cup \{e\}$ and $f \in I \cup \{e\}$.

In [1] it is shown that RIR-grammars generate exactly the context-free languages.

We will give another restricted from of indexed grammars which generate exactly the context-free languages.

Definition 2.6. An indexed grammar $G = (V, T, I, P, S)$ is called a rightlinear indexed grammar, if each production in P is of one of the forms $Af \to uB\gamma$ or $Af \to u$ with $A, B \in V$, $f \in I \cup \{e\}$, $u \in T^*$, and $\gamma \in I^*$.

Obviously, each RIR-grammar is a rightlinear indexed grammar. On the other hand, it is easy to show that for each rightlinear indexed grammar there is an equivalent RIR-grammar. Therefore we can state the following theorem.

Theorem 2.7. *Rightlinear indexed grammars generate exactly the context-free languages.*

If we define the analogous notion of a leftlinear indexed grammar, Theorem 2.7 holds if we replace rightlinear by leftlinear. Now it is easy to prove the following.

Theorem 2.8. *Let T be an alphabet, $c \notin T$ and $X, Y \subseteq T^*$. If $L = XcY$ is a linear indexed language, then X or Y is context-free.*

Remark. If we replace in this theorem 'linear indexed' by 'linear context-free' and 'context-free' by 'regular', we obtain a well-known theorem for linear context-free languages (see, e.g., [5, p. 222]).

Proof of Theorem 2.8. Let $G = (V, T \cup \{c\}, I, P, S)$ be a linear indexed grammar with $L(G) = XcY$. Define the following subsets of P:

$$P_0 = \{Af \to u \mid u \in T^*\} \cup \{Af \to uB\gamma v \mid u, v \in T^*,\ B\gamma \in VI^*\},$$

$$P_1 = \{Af \to ucv \mid u, v \in T^*\},$$

$$P_2 = \{Af \to u_1 c u_2 B\gamma v \mid u_1, u_2, v \in T^*,\ B\gamma \in VI^*\},$$

$$P_3 = \{Af \to uB\gamma v_1 c v_2 \mid u, v_1, v_2 \in T^*,\ B\gamma \in VI^*\}.$$

Set $G_i = (V, T \cup \{c\}, I, P_0 \cup P_i, S)$ and $L_i = L(G_i)$ for $i \in [1:3]$. It is easy to show that $L = L_1 \cup L_2 \cup L_3$ holds. Let now $S_1, S_2 : T^* c T^* \to T^*$ be mappings given by $S_1(ucv) = u$ and $S_2(ucv) = v$.

The rightlinear indexed grammar $G_1' = (V, T, I, P_1', S)$ with

$$P_1' = \{Af \to uB\gamma \mid Af \to uB\gamma v \in P_0\} \cup \{Af \to u \mid Af \to ucv \in P_1\}$$

generates $S_1(L_1)$. Using Theorem 2.7 we conclude that $S_1(L_1)$ is context-free.

The rightlinear indexed grammar $G_2' = (V', T, I, P_2', S)$ with $V' = V \cup \{A' \mid A \in V\}$ and

$$P_2' = \{Af \to uB\gamma \mid Af \to uB\gamma v \in P_0\}$$
$$\cup \{Af \to u_1 B'\gamma \mid Af \to u_1 c u_2 B\gamma v \in P_2\}$$
$$\cup \{A'f \to B'\gamma \mid Af \to uB\gamma v \in P_0\} \cup \{A'f \to e \mid Af \to u \in P_0\}$$

generates $S_1(L_2)$. With the same argument as above, $S_1(L_2)$ is context-free.

A similar construction of a leftlinear indexed grammar shows that $S_2(L_3)$ is context-free too.

It is now easy to show that $X = S_1(L_1) \cup S_1(L_2)$ or $Y = S_2(L_3)$ holds. This completes the proof. □

Now we can show the following theorem.

Theorem 2.9. *The class of linear indexed languages is not closed under product and Kleene closure.*

Proof. Let $L=\{ww \mid w\in\{a,b\}^*\}$ and $L'=Lc$. Obviously, L' is a linear indexed language. Since L is not context-free, together with Theorem 2.8 we conclude that $L'L$ is not a linear indexed language.

With the same argument, using properties of a full trio, L'^* is not a linear indexed language. □

3. Characterizations of linear indexed languages

In this section we will give two characterizations of linear indexed languages. Let us start with the first characterization which is based on controlling linear context-free grammars with context-free languages. The idea is to convert a linear context-free grammar with a context-free control language into a linear indexed grammar, where the indices are used to simulate leftmost derivations of the control words.

Theorem 3.1. *Let $C=(G,\mu,M,L)$ be a controlled linear grammar such that L is a context-free language. Then $L(C)$ is a linear indexed language.*

Proof. Let $G=(N,T,P,S)$ and let $G'=(N',M,S',P')$ be a context-free grammar in GNF (Greibach normal form) with $L=L(G')$.

Construct the linear indexed grammar $G_i=(V,T,I,P_i,S_i)$ with $V=N\cup\{S_i\}\cup\{F\}$, $I=N'\cup\{\#\}$, and the set of productions P_i is defined as follows:

(1) $S_i\to SS'\#\in P_i$.

(2) If $m: A\to uBv\in P$ and $A'\to mB'_1\dots B'_k\in P'$, then $AA'\to uBB'_1\dots B'_kv\in P_i$.

(3) If $m: A\to u\in P$ and $A'\to m\in P'$, then $AA'\to uF\in P_i$.

(4) $F\#\to e\in P_i$.

Here $A,B\in N$, $u,v\in T^*$, and $A',B'_1,\dots,B'_k\in N'$.

By an easy induction on n one can show: There is a $\pi\in M^*$ with $S\Rightarrow^n_\pi w_1Aw_2$ and $S'\Rightarrow^n\pi\alpha$ according to G and G' iff $S_i\Rightarrow SS'\#\Rightarrow^n w_1A\alpha\#w_2$ according to G_i.

Now let $w\in L(C)$. Hence there exist $\pi m\in M^*$, $m\in M$, with $S\Rightarrow_\pi w_1Aw_2\Rightarrow_m w_1uw_2=w$ according to G and $S'\Rightarrow^*\pi A'\Rightarrow\pi m$ according to G'. Since $m: A\to u\in P$ and $A'\to m\in P'$, we have $AA'\to uF\in P_i$, and together with the induction hypothesis we conclude that

$$S_i\Rightarrow SS'\#\Rightarrow^* w_1AA'\#w_2\Rightarrow w_1uF\#w_2\Rightarrow w_1uw_2=w$$

holds according to G_i. This shows the inclusion $L(C)\subseteq L(G_i)$.

Conversely, let $w\in L(G_i)$, i.e.,

$$S_i\Rightarrow SS'\#\Rightarrow^n w_1AA'\#w_2\Rightarrow w_1uF\#w_2\Rightarrow w_1uw_2=w$$

holds according to G_i. Therefore there exists $m\in M$ with $m: A\to u\in P$ and $A'\to m\in$

P', and with the aid of the induction hypothesis we conclude that

$$S \Rightarrow_{\pi}^{n} w_1 A w_2 \Rightarrow w_1 u w_2 = w \quad \text{and} \quad S' \Rightarrow^{n} \pi A' \Rightarrow \pi m$$

hold according to G and G' with a suitable $\pi \in M^*$. Hence $L(G_i) \subseteq L(C)$. □

We will now prove that each linear indexed language can be obtained by a linear context-free grammar controlled with a context-free language. To this end we need the following lemma.

Lemma 3.2. *Each linear indexed language can be generated by a linear indexed grammar with the property that in each derivation of a terminal word each introduced index will be consumed. Furthermore the left sides of all productions are of the form Af, $f \neq e$, except for the start production.*

Proof. Let $G_i = (V, T, I, P_i, S_i)$ be a linear indexed grammar. Construct a linear indexed grammar $G_i' = (V', T, I', P_i', S_i')$ with $V' = V \cup \{S_i', F\}$, S_i', F are new variables, $I' = I \cup \{\#\}$, $\#$ a new index, and the following set P_i' of productions:

(1) $S_i' \to S_i \# \in P_i'$.
(2) $Af \to uB\gamma v \in P_i'$ if $Af \to uB\gamma v \in P_i$ and $f \neq e$.
(3) $Ag \to uB\gamma g v \in P_i'$ for all $g \in I'$ if $A \to uB\gamma v \in P_i$.
(4) $Af \to uF \in P_i'$ if $Af \to u \in P_i$ and $f \neq e$.
(5) $Ag \to uFg \in P_i'$ for all $g \in I'$ if $A \to u \in P_i$.
(6) $Ff \to F \in P_i'$ for all $f \in I$.
(7) $F\# \to e \in P_i'$.

Here $A, B \in V$, $u, v \in T^*$, and $\gamma \in I^*$.

It is easy to see that $L(G_i) = L(G_i')$ holds. Furthermore G_i' has the above-mentioned properties. □

Now we can prove the following.

Theorem 3.3. *For each linear indexed grammar G_i there exists a controlled linear grammar $C = (G, \mu, M, L)$ such that L is a context-free language and $L(C) = L(G_i)$ holds.*

Proof. Let $G_i = (V, T, I, P_i, S_i)$ be a linear indexed grammar. Consider $G_i' = (V', T, I', P_i', S_i')$ constructed as in Lemma 3.2. Define a homomorphism $\phi : (V' \cup I' \cup T)^* \to (V' \cup T)^*$ by $\phi(A) = A$, $\phi(f) = e$ and $\phi(a) = a$ for all $A \in V'$, $f \in I'$ and $a \in T$.

Construct a linear context-free grammar $G = (N, T, P, S_i)$ with

$$N = V' - \{S_i'\} \quad \text{and} \quad P = \{A \to \phi(\alpha) \mid Af \to \alpha \in P_i' \text{ and } A \neq S_i'\}.$$

Choose an alphabet M and a surjective function $\mu : M \to P$, i.e., mark the productions of G with elements of M.

Construct a context-free grammar $G' = (N', M, S', P')$ with $N' = I'$, $S' = \#$ and the following set P' of productions:

(1) $f \to m\gamma \in P'$ if $m: A \to uBv \in P$ and $Af \to uB\gamma v \in P'_i$,

(2) $\# \to m$ if $m: F \to e \in P$.

Let L be the context-free language $L(G')$. Now we have constructed a controlled linear grammar $C = (G, \mu, M, L)$.

By a simple induction on n we can prove: $S'_i \Rightarrow S_i\# \Rightarrow^n uA\gamma\#v$ according to G'_i iff there exists a $\pi \in M^*$ with $|\pi| = n$ such that $S_i \Rightarrow_\pi uAv$ according to G and $\# \overset{n}{\Rightarrow} \pi\gamma\#$ according to G'. ($\overset{n}{\Rightarrow}$ denotes a leftmost derivation in n steps.) Here we have $u, v \in T^*$, $A \in V'$, $\gamma \in I'^*$ and $\pi \in M^*$. It is now easy to conclude $L(C) = L(G'_i) = L(G_i)$, which proves the theorem. □

Combining the two foregoing theorems, we have the following.

Theorem 3.4. *The linear indexed languages are exactly those languages which can be obtained by controlling linear context-free grammars with context-free languages.*

Remark. This theorem shows that the family of linear indexed languages coincides with the family $\mathscr{L}_1$ introduced by Khabbaz [7], which also implies that the family of context-free languages (the family $\mathscr{L}_0$ in [7]) is properly contained in the family of linear indexed languages. Furthermore, a special case of the pumping lemma in [7] yields a pumping lemma for linear indexed languages.

Now we will give the second characterization of linear indexed languages. It is known that the linear context-free languages are exactly the languages $\{\phi_1(w)\phi_2(w^R) \mid w \in U\}$, where U is a regular set and ϕ_1, ϕ_2 are homomorphisms (see [5, p. 64]). We will show that we arrive at the family of linear indexed languages if we substitute 'regular' for 'context-free'.

First we need the following theorem.

Theorem 3.5. *Let X, Y be alphabets, $\phi_1, \phi_2: X^* \to \mathbb{P}(Y^*)$ be context-free substitutions and let $L \subseteq X^*$ be a context-free language. Then $L_1 = \bigcup_{w \in L} \phi_1(w)\phi_2(w)^R$ is a linear indexed language.*

Proof. Let $G = (N, X, P, S)$ be a context-free grammar in GNF with $L = L(G)$. Furthermore, for all $a \in X$ let $G_i^a = (N_i^a, Y, P_i^a, S_i^a)$ be a context-free grammar in GNF with $L(G_i^a) = \phi_i(a)$ for $i \in [1:2]$. We can assume that the sets N, N_i^a, $a \in X$, $i \in [1:2]$ are pairwise disjoint.

We will now construct a linear indexed grammar which generates L_1. To this end set

$$G_1 = (\{\#\}, Y, I, P_1, \#) \quad \text{with } I = N \cup \bigcup_{a \in X} (N_1^a \cup N_2^a) \cup \{\not\in\}$$

(here $\#$ and $¢$ are new symbols) and define P_1 as follows:

(1) The productions $\# \to \# S¢$ and $\#¢ \to e$ are in P_1.

(2) If $S \to e \in P$, then $\# S \to \# \in P_1$.

(3) For each production $A \to aB_1 \ldots B_k \in P$ with $a \in X$, $k \geq 0$, the production $\# A \to \# S_1^a S_2^a B_1 \ldots B_k$ is in P_1.

(With the aid of these productions, derivations of words $w \in L$ are simulated in the index words following $\#$.)

(4) If $S_i^a \to e \in P_i^a$, then $\# S_i^a \to \# \in P_1$ for $a \in X$, $i \in [1:2]$.

(5) If $D \to bE_1 \ldots E_k \in P_1^a$, where $a \in X$, then $\# D \to b\# E_1 \ldots E_k \in P_1$ with $b \in Y$ and $k \geq 0$.

(6) If $D \to bE_1 \ldots E_k \in P_2^a$ where $a \in X$, then $\# D \to \# E_1 \ldots E_k b \in P_1$ with $b \in Y$ and $k \geq 0$.

(With the aid of (4), (5), derivations of words $u \in \phi_1(a)$, $a \in X$ are simulated in the index words following $\#$. The same holds for derivations of words $v \in \phi_2(a)$, $a \in X$, w.r.t. (4), (6).)

Obviously G_1 is a linear indexed grammar.

To prove $L(G_1) = L_1$, we will first show the following three assertions.

(i) $\phi_1(a) = \{u \mid u \in Y^* \text{ and } \# S_1^a \Rightarrow^* u\#\}$, and
$\phi_2(a) = \{v \mid v \in Y^* \text{ and } \# S_2^a \Rightarrow^* \# v^R\}$ for all $a \in X$.

(ii) $\phi_1(w)\phi_2(w)^R \subseteq L(G_1)$ for all $w \in L$.

(iii) For all $u \in L(G_1)$ there exists a $w \in L$ such that $u \in \phi_1(w)\phi_2(w)^R$.

By easy inductions on n one can prove:

- $D \Rightarrow^n w$ holds according to G_1^a iff $\# D \Rightarrow^n w\#$ holds according to G_1, where $w \in Y^*$, and
- $D \Rightarrow^n w$ holds according to G_2^a iff $\# D \Rightarrow^n \# w^R$ holds according to G_1, where $w \in Y^*$.

This shows (i).

If $A \Rightarrow^n w$, $w \in X^*$, holds according to G, then $\# A \Rightarrow^* v_1 \# v_2^R$ holds according to G_1 for all $v_1 \in \phi_1(w)$ and $v_2 \in \phi_2(w)$. This together with (1) of the definition of P_1 proves (ii).

If $\# A \Rightarrow^n u\# v$ holds according to G_1, where $A \in N$ and $u, v \in Y^*$, then there exists a $w \in X^*$ with $u \in \phi_1(w)$, $v \in \phi_2(w)^R$ and $A \Rightarrow^* w$ holds according to G.

This together with (1) of the definition of P_1 proves (iii).

From (ii) the inclusion $L_1 \subseteq L(G_1)$ follows. The inverse inclusion follows from (iii). This proves the theorem. □

Corollary 3.6. *Let $L \subseteq X^*$ be a context-free language and let $h_1, h_2 : X^* \to Y^*$ be homomorphisms. Then $L_1 = \{h_1(w)h_2(w)^R \mid w \in L\}$ is a linear indexed language.*

Corollary 3.7. *Let $L \subseteq X^*$ be a context-free language. Then $L_1 = \{ww^R \mid w \in L\}$ is a linear indexed language.*

Let us now give our second characterization of linear indexed languages.

Theorem 3.8. *If $L_1 \subseteq Y^*$ is a linear indexed language, then there exists an alphabet X, a context-free language $L \subseteq X^*$ and two homomorphisms $h_1, h_2: X^* \to Y^*$ such that $L_1 = \{h_1(w)h_2(w)^R \mid w \in L\}$ holds.*

Proof. Let $G_1 = (V, Y, I, P_1, S)$ be a linear indexed grammar with $L_1 = L(G_1)$. Let $p_1, p_2, \ldots, p_k$ be the productions in P. Set $X = \{p_1, \ldots, p_k\}$ and define a rightlinear indexed grammar $G = (V, X, I, P, S)$ and homomorphisms $h_1, h_2: X^* \to Y^*$ in the following way:

(a) If $p_i: Af \to uB\gamma v$, then $Af \to p_i B\gamma \in P$, $h_1(p_i) = u$ and $h_2(p_i) = v^R$.

(b) If $p_i: Af \to u$, then $Af \to p_i \in P$, $h_1(p_i) = u$ and $h_2(p_i) = e$.

$L = L(G)$ is context-free since G is a rightlinear indexed grammar.

Obviously we have the following connections between derivations according to G_1 and G: If $S \Rightarrow^* \pi$ holds according to G, then $S \Rightarrow_\pi h_1(\pi)h_2(\pi)^R$ holds according to G_1.

If $S \Rightarrow^* v$, $v \in Y^*$, holds according to G_1, then there exists a $\pi \in X^*$ with $h_1(\pi)h_2(\pi)^R = v$ and $S \Rightarrow^* \pi$ holds according to G.

Therefore, we conclude $L_1 = \{h_1(w)h_2(w)^R \mid w \in L\}$, which proves the theorem. □

Combining Theorems 3.5 and 3.8 we arrive at our second characterization of linear indexed languages.

Theorem 3.9. *A language $L_1 \subseteq Y^*$ is a linear indexed language iff there exists an alphabet X, a context-free language $L \subseteq X^*$ and two homomorphisms $h_1, h_2: X^* \to Y^*$, such that $L_1 = \{h_1(w)h_2(w)^R \mid w \in L\}$ holds.*

This characterization yields the following closure property of linear indexed languages.

Theorem 3.10. *The class of linear indexed languages is closed under context-free substitutions.*

4. A generator for linear indexed languages

As mentioned in Section 2, the family of linear context-free languages and linear indexed languages are full trios. Furthermore, it is known that the family of linear context-free languages is generated by any symmetric language S_n, $n \geq 2$. Here S_n is the linear context-free language generated by the linear context-free grammar $G_n = (\{S\}, X_n, P_n, S)$, where $X_n = \{x_1, x_1', \ldots, x_n, x_n'\}$ is a paired alphabet and $P_n = \{S \to x_i S x_i' \mid i \in [1:n]\} \cup \{S \to e\}$. This implies, that the family of linear context-free languages is a full principal semi-AFL (for these notions, see, e.g., [2] or [4]).

In this section we will prove that the family of linear indexed languages is a full principal semi-AFL too. Let us first outline the idea of the proof. We already know that the linear indexed languages are exactly those languages which can be obtained by controlling linear context-free grammars with context-free languages (Theorem 3.4).

Now let L_1 be a linear indexed language. There exists a controlled linear grammar $C=(G, \mu, M, L)$, L context-free, with $L_1=L(C)$. For each $w_1 \in L_1$ there exists a word $m_1 \dots m_k m \in L$ and a derivation

$$S \Rightarrow_{m_1} u_1 A_1 v_1 \Rightarrow_{m_2} u_1 u_2 A_2 v_2 v_1 \Rightarrow \cdots$$

$$\Rightarrow_{m_k} u_1 u_2 \dots u_k A_k v_k \dots v_2 v_1 \Rightarrow_m u_1 \dots u_k u v_k \dots v_1 = w_1 \qquad (\bigstar)$$

according to G. Therefore we can obtain w_1 in the following way: According to $\pi = m_1 \dots m_k m \in L$ we generate the prefix $u_1 \dots u_k u$ of w_1 and then, according to $\pi^R = m m_k \dots m_1$ we generate the suffix $v_k \dots v_1$ of w_1.

From the Chomsky–Schützenberger Theorem for context-free languages we know that, given L, there exist $n \geq 2$, a regular language U and a homomorphism ϕ such that $L=\phi(D_n \cap U)$ holds. Here D_n is the semi-Dyck language over the paired alphabet $X_n=\{x_1, x_1', \dots, x_n, x_n'\}$. Therefore $\pi \in L$ can be given as $\phi(w)$, $w \in D_n \cap U$. Moreover, ϕ can be chosen as an alphabetic homomorphism, i.e., $|\phi(x)| \leq 1$ for all $x \in X_n$.

Now consider the language $D_{in}=\{w \#^2 w^R \mid w \in D_n\}$, where $\# \notin X_n$. From Corollary 3.7 we know that D_{in} is a linear indexed language.

We will construct a gsm T with $T(D_{in})=L_1$, which operates as follows. (For the definition of a gsm, see Hopcroft and Ullman [6, p. 272]. A gsm is an a-transducer in the sense of Ginsburg [4].) If $w \#^2 w^R \in D_{in}$ with $w \in U$ is entered in T such that $(\bigstar)$ holds for $\pi=\phi(w)$, then, after reading w, the output of T is $u_1 \dots u_k u$, and then, reading $\#^2 w^R$, T writes $v_k \dots v_1$.

Then, using the fact that each D_{in}, $n \geq 2$, is the image of D_{12} by a gsm-mapping, we arrive at our result.

Theorem 4.1. *For each linear indexed language L_1 there exist $n \geq 2$ and a gsm T such that $L_1=T(D_{in})$ holds.*

Proof. Let $L_1 \subseteq Y^*$ be a linear indexed language. By Theorem 3.4 there exists a controlled linear grammar $C=(G, \mu, M, L)$, where $G=(N, Y, P, S)$ is a linear context-free grammar and L a context-free language such that $L_1=L(C)$ holds. Since $L \subseteq M^*$ is context-free, there exist $n \geq 2$, a regular language $U \subseteq X_n^*$ and an alphabetic homomorphism $\phi: X_n^* \to M^*$ with $L=\phi(D_n \cap U)$, where $X_n=\{x_1, x_1', \dots, x_n, x_n'\}$ is a paired alphabet and D_n is the semi-Dyck language over X_n.

Let $A=(Q', X_n, \delta', q_0', F')$ be a deterministic finite state acceptor for U. Construct the gsm $T=(Q, X_n \cup \{\#\}, Y, \delta, q_0, F)$ with $Q=Q' \times (N \cup \{e\}) \cup \{r, p\}$, $\# \notin X_n$, $q_0=[q_0', S]$, $F=\{p\}$ and δ is defined as follows:

(1) For all $q \in Q'$, $x \in X_n$:

(1.1) If $\phi(x) = e$, then $\delta([q, A], x) = ([\delta'(q, x), A], e)$ for all $A \in N \cup \{e\}$. Furthermore, set $\delta(p, x) = (p, e)$.

(1.2) If $\phi(x) = m : A \to uBv \in P$, then $\delta([q, A], x) = ([\delta'(q, x), B], u)$ and $\delta(p, x) = (p, v)$.

(1.3) If $\phi(x) = m : A \to u \in P$, then $\delta([q, A], x) = ([\delta'(q, x), e], u)$ and $\delta(p, x) = (p, e)$.

(2) $\delta([q, e], \#) = (r, e)$ for all $q \in F'$ and $\delta(r, \#) = (p, e)$.

Let us first state some properties of T. We assume that the transition function δ' of A is extended to $Q' \times X_n^*$ in the usual way.

Claim 1. *If* $([q, A], w, e) \vdash^* ([q', B], e, u)$ *holds with* $A, B \in N \cup \{e\}$, $w \in X_n^*$, *then* $q' = \delta'(q, w)$.

Claim 2. *If* $([q_0', S], w, e) \vdash^* ([q, A], e, u)$ *holds for* $w \in X_n^*$, $u \in Y^*$, $A \in N \cup \{e\}$ *and* $q \in F'$, *then* $w \in U$. *If furthermore* $w \in D_n$, *then* $\phi(w) \in L$.

Claim. 3. *If* $w \in X_n^*$ *with* $\phi(w) = e$, *then* $([q, A], w, e) \vdash^* ([\delta'(q, w), A], e, e)$ *holds for all* $A \in N \cup \{e\}$ *and* $q \in Q'$. *Therefore,* $([q, A], w, e) \vdash^* ([\delta'(q, w), e], e, u)$ *with* $A \in N$ *implies* $\phi(w) \neq e$.

Claim 4. *If* $([q_0', S], w, e) \vdash^* ([q, A], e, u)$ *for* $A \in N \cup \{e\}$, $w \in X_n^*$, *then* $S \Rightarrow_\pi uAv$ *with* $\pi = \phi(w)$ *and* $(p, w^R, e) \vdash^* (p, e, v)$.

Proof. Let $|w| = 0$, i.e., $w = e$ and $\phi(w) = e$. Then $A = S$, and we have $S \Rightarrow_e S$ and $(p, e, e) \vdash^* (p, e, e)$.

Let $w = w'x$, $x \in X_n$.

(a) If $\phi(x) = e$, then

$$([q_0', S], w'x, e) \vdash^* ([q, A], x, u) \vdash ([\delta'(q, x), A], e, u).$$

From the induction hypothesis we have $S \Rightarrow_\pi uAv$ with $\pi = \phi(w')$ and $(p, w'^R, e) \vdash^* (p, e, v)$. From $\phi(w') = \phi(w) = \pi$ and $(p, xw'^R, e) \vdash^* (p, e, v)$ the assertion follows.

(b) If $\phi(x) = m$: $A \to u_1 B v_1$, $B \in N$ or $B = e$ and $v_1 = e$, then

$$([q_0', S], w'x, e) \vdash^* ([q, A], x, u) \vdash ([\delta'(q, x), B], e, uu_1).$$

From the induction hypothesis we have $S \Rightarrow_{\pi'} uAv$ with $\pi' = \phi(w')$ and $(p, w'^R, e) \vdash^* (p, e, v)$.

Therefore, we have $S \Rightarrow_\pi uu_1 B v_1 v$ with $\pi = \pi' m$ and $(p, xw'^R, e) \vdash^* (p, e, v_1 v)$.

Claim 5. *If* $S \Rightarrow_\pi uAv$ *with* $A \in N$, *then for all* $w \in X_n^*$ *with* $\phi(w) = \pi$ *we have* $([q_0', S], w, e) \vdash^* ([q, A], e, u)$ *and* $(p, w^R, e) \vdash^* (p, e, v)$.

Proof. Let $|\pi| = 0$ and let $\phi(w) = \pi$. Then $u = v = e$ and $A = S$. From Claim 3 we have $([q_0', S], w, e) \vdash^* ([\delta'(q_0', w), S], e, e)$. Furthermore $(p, w^R, e) \vdash^* (p, e, e)$ holds.

Let $\pi = \pi' m$ and let $w \in X_n^*$ with $\phi(w) = \pi$. Then $w = w'xw''$ with $\phi(w') = \pi'$, $\phi(x) = m$ and $\phi(w'') = e$. We have $S \Rightarrow_{\pi'} u'A'v' \Rightarrow_m u'u''Av''v' = uAv$.

With the aid of the induction hypothesis we conclude

$$([q_0', S], w'xw'', e) \vdash^* ([q', A'], xw'', u') \vdash ([q'', A], w'', u'u'') \vdash^* ([q, A], e, u'u'').$$

Furthermore,

$$(p, w^R, e) = (p, w''^R x w'^R, e) \vdash^* (p, x w'^R, e) \vdash (p, w'^R, v'') \vdash^* (p, e, v''v')$$
$$= (p, e, v).$$

Now let $w_1 \in T(D_{in})$. Then there exists a $w \in D_n$ with $([q_0', S], w\#^2 w^R, e) \vdash^* (p, e, w_1)$. Therefore

$$([q_0', S], w\#^2 w^R, e) \vdash^* ([q, e], \#^2 w^R, u) \vdash (r, \# w^R, u) \vdash (p, w^R, u)$$

with $q \in F'$. Then $w \in U$ (see Claim 2), which implies $\phi(w) \in L$. From Claim 4 we conclude $S \Rightarrow_\pi uv$ with $\pi = \phi(w)$ and $(p, w^R, e) \vdash^* (p, e, v)$. Therefore, $w_1 = uv \in L_1$. This shows the inclusion $T(D_{in}) \subseteq L_1$.

Now let $w_1 \in L_1$, i.e., $S \Rightarrow_\pi uAv \Rightarrow_m uu'v$ with $\pi m \in L$. Then there exists a $w \in D_n \cap U$ with $\phi(w) = \pi m$. Let $w = w'xw''$ with $\phi(w') = \pi$, $\phi(x) = m$ and $\phi(w'') = e$. From Claim 5 we have

$$([q_0', S], w\#^2 w^R, e) \vdash^* ([q', A], xw''\#^2 w^R, u) \quad \text{and} \quad (p, w'^R, e) \vdash^* (p, e, v).$$

Since $\phi(x) = m : A \to u'$, we have

$$([q', A], xw''\#^2 w^R, u) \vdash ([q'', e], w''\#^2 w^R, uu').$$

From Claim 3,

$$([q'', e], w''\#^2 w^R, uu') \vdash^* ([q, e], \#^2 w^R, uu')$$

follows. From Claim 1 we have $\delta'(q_0', w) = q$, and since $w \in U$, this implies $q \in F'$.

Hence, we can proceed in the following way:

$$([q, e], \#^2 w^R, uu') \vdash^2 (p, w''^R x w'^R, uu') \vdash^* (p, x w'^R, uu')$$
$$\vdash (p, w'^R, uu') \vdash^* (p, e, uu'v).$$

Therefore, we have $L_1 \subseteq T(D_{in})$. This completes the proof. □

It remains to show that each D_{in}, $n \geq 2$, is the image of D_{i2} under a gsm-mapping. Using well-known proof techniques, the following theorem can easily be shown.

Theorem 4.2. *For each D_{in}, $n \geq 2$, there is a homomorphism ϕ such that $D_{in} = \phi^{-1}(D_{i2})$ holds.*

Combining Theorems 4.1 and 4.2 we arrive at the main result of this section.

Theorem 4.3. *The family of linear indexed languages is a full principal semi AFL.*

5. Parikh mappings of linear indexed languages

The characterization of linear indexed languages given in Theorem 3.9 allows us to state a Parikh theorem for linear indexed languages. Since there are indexed

languages which yield no semilinear set under a Parikh mapping we can conclude that the linear indexed languages form a proper subclass of the indexed languages. For the definition of a Parikh mapping and a semilinear set, see, e.g., [5, p. 226]. Let $X=\{x_1,\ldots,x_n\}$ and $Y=\{y_1,\ldots,y_m\}$ be alphabets and $h_1, h_2: X^* \to Y^*$ be homomorphisms. Furthermore let $\psi: X^* \to \mathbb{N}_n$ and $\psi_1: Y^* \to \mathbb{N}_m$ be Parikh mappings.

Define $\Phi: \mathbb{N}_n \to \mathbb{N}_m$ as $\Phi(\alpha)=\psi_1(h_1(w)h_2(w))$, where $\alpha \in \mathbb{N}_n$ and $w \in X^*$ with $\psi(w)=\alpha$. Obviously, Φ is a well-defined function. Furthermore, Φ is a homomorphism. To this end take $\alpha, \beta \in \mathbb{N}_n$ and let $\alpha=\psi(w_1)$ and $\beta=\psi(w_2)$ with $w_1, w_2 \in X^*$. Then $\psi(w_1w_2)=\alpha+\beta$ and

$$\begin{aligned}\Phi(\alpha+\beta) &= \psi_1(h_1(w_1w_2)h_2(w_1w_2)) \\ &= \psi_1(h_1(w_1)h_1(w_2)h_2(w_1)h_2(w_2)) \\ &= \psi_1(h_1(w_1)h_2(w_1)) + \psi_1(h_1(w_2)h_2(w_2)) \\ &= \Phi(\alpha)+\Phi(\beta).\end{aligned}$$

Theorem 5.1. *Let $Y=\{y_1,\ldots,y_m\}$ be an alphabet and $L_1 \subseteq Y^*$ be a linear indexed language. Then $\psi_1(L_1)$ is a semilinear set, where $\psi_1: Y^* \to \mathbb{N}_m$ is the Parikh mapping.*

Proof. According to Theorem 3.9 there exist an alphabet $X=\{x_1,\ldots,x_n\}$, a context-free language $L \subseteq X^*$ and two homomorphisms $h_1, h_2: X^* \to Y^*$ such that $L_1=\{h_1(w)h_2(w)^R \mid w \in L\}$ holds. Let $\psi: X^* \to \mathbb{N}$ be the Parikh mapping. Then $\psi(L)$ is a semilinear set, i.e., there are linear sets $M_1,\ldots,M_r$, $r \geq 0$, such that $\psi(L)=\bigcup_{i=1}^{r} M_i$.

Let Φ be defined as above. Then $\Phi(M_i)=R_i$ are linear sets since Φ is a homomorphism. Now let $v=h_1(w)h_2(w)^R \in L_1$, where $w \in L$. Then $\psi(w) \in M_i$ for an $i \in [1:r]$, and

$$\Phi(\psi(w))=\psi_1(h_1(w)h_2(w))=\psi_1(h_1(w)h_2(w)^R)=\psi_1(v) \in R_i.$$

Therefore $\psi_1(L_1) \subseteq \bigcup_{i=1}^{r} R_i$.

Now let $\beta \in R_i$ for an $i \in [1:r]$. Then there exists an $\alpha \in M_i$ with $\Phi(\alpha)=\beta$. Furthermore, there exists a $w \in L$ with $\alpha=\psi(w)$. Since

$$h_1(w)h_2(w)^R \in L_1 \quad \text{and} \quad \psi_1(h_1(w)h_2(w)^R)=\psi_1(h_1(w)h_2(w))=\Phi(\alpha)=\beta$$

we conclude $\bigcup_{i=1}^{r} R_i \subseteq \psi_1(L_1)$.

This completes the proof of the theorem. □

Recall that two languages are called letter-equivalent if they have the same image under the Parikh mapping.

Corollary 5.2. *Each linear indexed language is letter-equivalent to a regular set.*

Now we can prove the following inclusion theorem.

Theorem 5.3. *The family of linear indexed languages is a proper subclass of the family of indexed languages.*

Proof. The language $\{a^{2^k} \mid k \geq 1\}$ is an indexed language but not a regular set. □

Acknowledgment

The authors wish to thank the referee for his (or her) useful hints and suggestions.

References

[1] A.V. Aho, Indexed grammars, *J. ACM* **15** (1968) 647–671.
[2] J. Berstel, *Transductions and Context-Free Languages* (Teubner, Stuttgart, 1979).
[3] S. Ginsburg and E.H. Spanier, Control sets on grammars, *Math. Systems Theory* **2** (1968) 169–177.
[4] S. Ginsburg, *Algebraic and Automata-Theoretic Properties of Formal Languages* (North-Holland, Amsterdam, 1975).
[5] M.A. Harrison, *Introduction to Formal Language Theory* (Addison-Wesley, Reading, MA, 1978).
[6] J.E. Hopcroft and J.D. Ullman, *Introduction to Automata Theory, Languages and Computation* (Addison-Wesley, Reading, MA, 1979).
[7] N.A. Khabbaz, A geometric hierarchy of languages, *J. Comput. System. Sci.* **8** (1974) 142–157.
[8] R. Parchmann, Balanced context-free languages and indexed languages, *Elektron. Informationsverarbeit. u. Kybernetik*, to appear.
[9] A. Salomaa, *Formal Languages* (Academic Press, New York, 1973).